Defence Research and Development Canada
Recherche et développement pour la défense Canada

DEFENCE R&D DÉFENSE

Iterative sub-setting

Etienne Martineau
DRDC Valcartier

Defence R&D Canada – Valcartier
Technical Memorandum
DRDC Valcartier TM 2010-461
December 2010

Canada

Iterative sub-setting

Etienne Martineau
DRDC Valcartier

Defence R&D Canada – Valcartier

Technical Memorandum
DRDC Valcartier TM 2010-461
December 2010

Principal Author

Original signed by Etienne Martineau

Etienne Martineau

Defense Scientist

Approved by

Original signed by Stéphane Paradis

Stéphane Paradis

Section Head / Intelligence & Information Section, DRDC Valcartier

Approved for release by

Original signed by Christian Carrier

Christian Carrier

Chief Scientist, DRDC Valcartier

DRDC project 11hk

Abstract

The objective of this document is to communicate the progress made on an initiative to improve the current Case-Based Reasoning (CBR) functionality of the Multi-Intelligence Tools Suite (MITS) platform. To do so, a new similarity measure based on information theory called Iterative Sub-Setting (ISS) is proposed. The goal of this initiative was to provide a potential solution to a list of issues identified in the literature about Maritime Anomaly Detection (MAD). Various tests with this measure were made to demonstrate the potential for it to address selected issues in MAD. Results show that this method is more sensible to the context where the CBR is performed, can be used with a small case base, evolve over time and provide an accurate confidence level on the selected cases.

Résumé

L'objectif de ce document est de communiquer les progrès réalisés grâce à une initiative pour améliorer la capacité actuelle de raisonnement à base de cas de la plate-forme « Multi-Intelligence Tools Suite (MITS) ». Le but de cette initiative était de fournir une solution potentielle à une liste de problèmes identifiées dans la littérature sur la détection d`anomalies maritimes (DAM). On a proposé une nouvelle mesure de similarité basée sur la théorie de l'information. Différents tests ont été effectués avec cette mesure pour démontrer son potentiel à résoudre certaines problèmes en DAM. Les résultats montrent que cette méthode est plus sensible au contexte dans lequel le raisonnement à base de cas est effectué; elle peut être utilisée avec une base de cas limitée, évoluer dans le temps et fournir un niveau de confiance précis sur la pertinence des cas sélectionnés.

This page intentionally left blank.

Executive summary

Iterative sub-setting

E. Martineau; DRDC Valcartier TM 2010-461; Defence R&D Canada – Valcartier; December 2010.

Introduction or background: This Technical Memorandum documents some results and findings arising from research activities conducted at Defence R&D Canada (DRDC) under project 11hk, « Multiple Hypothesis Link Analysis for Anomaly Detection in the Maritime Domain », which is falling under DRDC's Applied Research Program (ARP). The objective of this document is to communicate the progress made on an initiative to improve the current Case-Based Reasoning (CBR) functionality of the Multi-Intelligence Tools Suite (MITS) platform. To do so, a new similarity measure based on information theory called Iterative Sub-Setting (ISS) is proposed.

Results: Results show that this method is more sensible to the context where the CBR is performed, can be used with a small case base, evolve over time and provide an accurate confidence level on the selected cases.

Significance: The goal of this initiative was to provide a potential solution to a list of issues identified in the literature about Maritime Anomaly Detection (MAD). The proposed initiative improves the capacity of detection of anomalous behaviour in the maritime domain of the current MITS platform.

Future plans: The results and findings reported in this memorandum constitute exploratory work. The reported research effort should be considered as a starting point for a more complete investigation of ISS potential. Efforts need to be invested to evaluate potentially valuable improvements ISS may procure.

Sommaire

Iterative sub-setting

E. Martineau; DRDC Valcartier TM 2010-461; R & D pour la défense Canada – Valcartier; Décembre 2010.

Introduction ou contexte: Ce mémorandum technique documente certains résultats et conclusions découlant des activités de recherche menées à R & D pour la Défense Canada (RDDC) au sein du projet 11hk, «Multiple Hypothesis Link Analysis for Anomaly Detection in the Maritime Domain», dans le cadre du Programme de recherche appliquée (ARP) de RDDC. L'objectif de ce document est de communiquer les progrès réalisés dans une initiative visant à améliorer la capacité de raisonnement à base de cas (CBR) de la plate-forme de la suite d'outils multirenseignements (MITS). Pour ce faire, une nouvelle mesure de similarité basée sur la théorie de l'information appelée Itérative Sub-Setting (ISS) est proposée.

Résultats: Les résultats montrent que cette méthode est plus sensible au contexte où le CBR est effectué. Elle peut être utilisée avec une base de cas réduite, évoluer au fil du temps et fournir un niveau de confiance sur la sélection des cas.

Importance: L'objectif de cette initiative était de fournir une solution potentielle à la liste des problèmes recensés dans la littérature sur la détection des anomalies maritimes (MAD). L'initiative proposée améliore la capacité de détection des comportements anormaux dans le domaine maritime de la plate-forme MITS actuelle.

Perspectives: Les résultats et conclusions présentés dans ce mémorandum constituent un travail exploratoire. Cet effort de recherche doit être considéré comme un point de départ pour une enquête plus complète sur le potentiel de l'ISS. Des efforts doivent être investis dans l'évaluation des améliorations potentiellement utiles que l'ISS peut procurer.

Table of contents

List of figures

1. Introduction

The objective of this document is to communicate the progress made on an initiative to improve the current Case-Based Reasoning (CBR) functionality for the Multi-Intelligence Tools Suite (MITS) platform. The work presented here is at an early stage and contains no actual improvement to the MITS. The document describes the motivation of this work, a new way to perform case-based reasoning and some empirical results applicable to a future version of the MITS. This effort was made in the context of Maritime Anomaly Detection (MAD). However, the results, like the CBR in the MITS, are domain agnostic.

The goal of this initiative was to provide a potential solution to a list of issues identified in the literature about MAD. Efforts were put to provide an overview of the potential benefits and risks of the new method. By doing so, minimal efforts were dedicated to provide decisional information on whatever research on this avenue shall continue.

Proof, details and formulas have been left aside to provide the reader with the general idea of this work. Moreover, only minimal efforts have been made to explain the research context; the aim of this document is not to provide a literature review. References to other works are provided.

2. Motives for this work

After much reading of different works in the field of maritime anomaly detection, it was clear that this field was full of opportunities. There is a lot to be done and, so far, most of the work has been to retrofit previous research in anomaly detection. The results of these works have shown some mitigated success. However, while they have not been up to the expectations, they have enlightened researchers on where to put new efforts. Recent talks and workshops have identified gaps in capabilities that need research efforts. This section provides the main motive of the work reported in this document, i.e. problems, gaps and new ideas.

2.1. Maritime domain anomaly detection

Recently, the domain of maritime anomaly detection has gained a lot of attention. Many researchers focus their efforts in finding outliers that could represent a potential threat. While anomaly detection is a well known field, its application to the maritime domain has not given good results so far. In fact, anomaly detection techniques give the results one could expect. They give a simplified model to explain the majority of the data. The unexplained part of the data is considered as an anomaly. The needs of the maritime domain are to declare as an anomaly any unexplainable behaviours in regards of all the data. These two definitions may seem identical but there is a critical difference.

Since the models are simplified, they do not explain everything. Anomalies are declared while the models don't take into account the appropriate evidence to explain a normal behaviour. One can define two kinds of anomalies: the real unexplainable ones and the failure of the modeling techniques. Usually no distinction is made between these two types of anomaly. They are only globally qualified under the labels "false positive" and "true negative". Here are some papers where this problem is present: [Bomberger et al, 2006];[Guerriero et al, 2010];[Rhodes et al, 2007];[Riveiro et al, 2008].

That said, operators of maritime surveillance systems don't care about the underlying models. They trust the system and they expect it to exploit every bit of information to assess the situation. So far, almost all research in the field provides solutions that encompass only a portion of the problem and let operators believe that everything is under control. An operator should know when a classification is done with an inappropriate model. The model should handle every situation, justify every decision and provide a confidence level.

2.2. Anomaly in context

The context in which events occur can help explain them. A particular event can only be tagged as an anomaly if the context cannot justify it. For most anomaly detection methods, the context is limited to a collection of similar events where a particular observation appears as an outlier. To put an event in context, one must find a relevant subset in this collection where the event can appear as normal or abnormal. Another way to call this practice is "Putting things in perspective".

An easy way to achieve this is to add relevant contextual data to an observation. By doing so, one raises the dimensionality of a dataset. For example, adding the gender, the age, the nationality and the religion to the profile of an individual can help situate him/her in his/her context. In contrast, some methods are focused on reducing the dimensionality by removing what is mostly believed as irrelevant data using methods like Principal Component Analysis (PCA). In this case, some information is lost and the problem discussed in the previous section can occur.

2.2.1. Issues with time

Handling time in maritime anomaly detection is a serious problem. To picture the problem, let's give a simple example. Let's imagine a river where hundreds of cargo ships travel up and down on it every day. Once every week, on Sunday afternoon, a ferry crosses the river. This ferry will be flagged as an anomaly by most methods because this event is too sporadic and will be considered like noise. However, it is clearly a pattern and a model should handle it. Figure 1 pictures this example taken from [Seibert, 2009]. Models should be able to classify patterns in space and time. Actually, vessel tracks are considered as time series and it is almost the only place where time is taken into account.

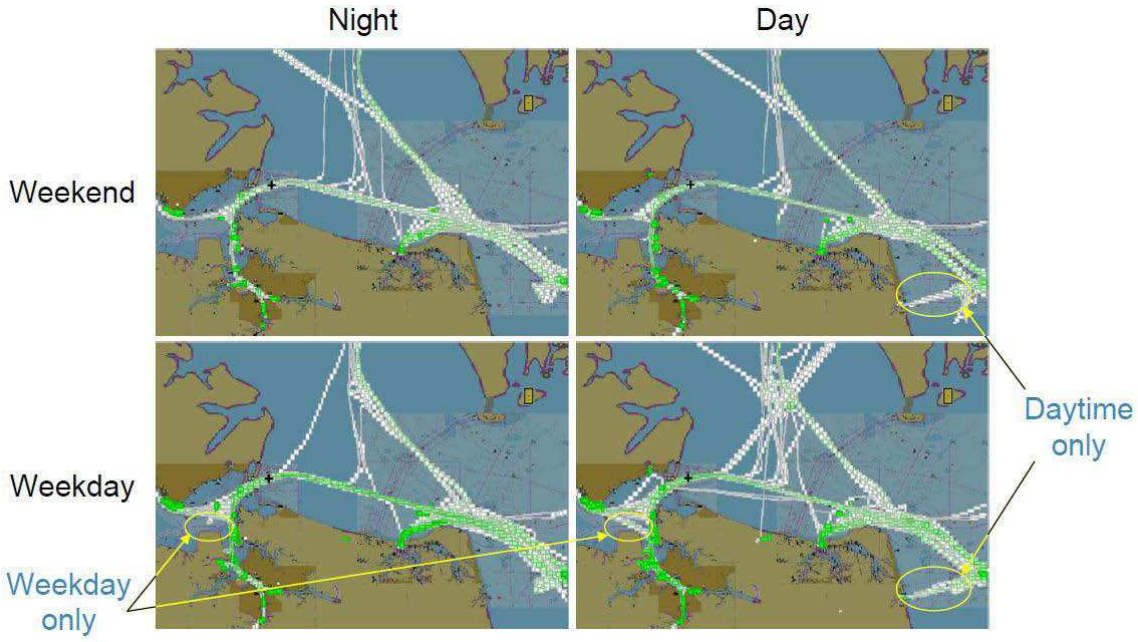

Figure 1 Pattern in time [Seibert, 2009]

2.2.2. Evolving models

Models are trained with historical data and then put online for classification. Most of the time, models are static and are subject to become obsolete [Berry, Linoff, 1997]. In few cases, the model can be dynamic. But again, models almost never evolve to follow the reality without human interaction. An obsolete model raises unjustified anomalies and decreases the confidence of the operator about it. Models should evolve with time and experience to be useful to the operator.

2.2.3. Hierarchical context

Levels of abstraction and generalization are useful to the cognitive process. Anomaly detection using hierarchical contexts give the opportunity to analyze the data under different perspectives. In fact, the information a system possesses may not be relevant at the vessel level but it may be related to families of vessels. Assessment of the behaviour of a vessel can be done for example using country of origin (like Canada). This behaviour may not be specific to Canadian vessels and can be typical of North-American vessels. In data-mining, the process of generalization is called "roll-up" and the process of specification is called "drill-down" [Han, Kambert, 2001]. These concepts are almost never used in maritime anomaly detection.

2.3. Supporting the decision process

System operators are often under a lot of pressure; they must do more with less. While the number of vessels is increasing, the staffing is being reduced [DARPA, 2005]. The operators must take justified decisions and assume the responsibility of the consequences. Anomaly detection systems are their tools and they rely on them to provide the maximum of information about a situation. Information-oriented anomaly detection methods are almost inexistent. Basics tools to achieve this are well known but appear to be left aside.

In fact, in current applications, an event is declared an anomaly based on its distance from a subset of data. The information content of this data can be totally irrelevant. It is important for the decision maker to understand why an anomaly is raised. That is why the reference subset must be chosen carefully. From the decision maker perspective, it does not help much if the system says "cars are bad because apples are bad". It has been shown that while anomaly detection is helpful, most systems have issues with the usability of the alerts an anomaly rises [Hutchins et al, 2009].

2.3.1. Information in missing values

Popular anomaly detection methods cannot handle missing values and, unfortunately, missing values are common. Common methods to handle missing values are to perform imputation or simply delete an entry. In an anomaly detection system, one cannot ignore an event because some fields are missing. All events must be handled and deletion is not an option. Imputation is problematic, since it assigns the most probable estimation and, by doing so, can possibly reduce the information content. The reason for this is that information is given by : -ln(x) as shown in Figure 2.

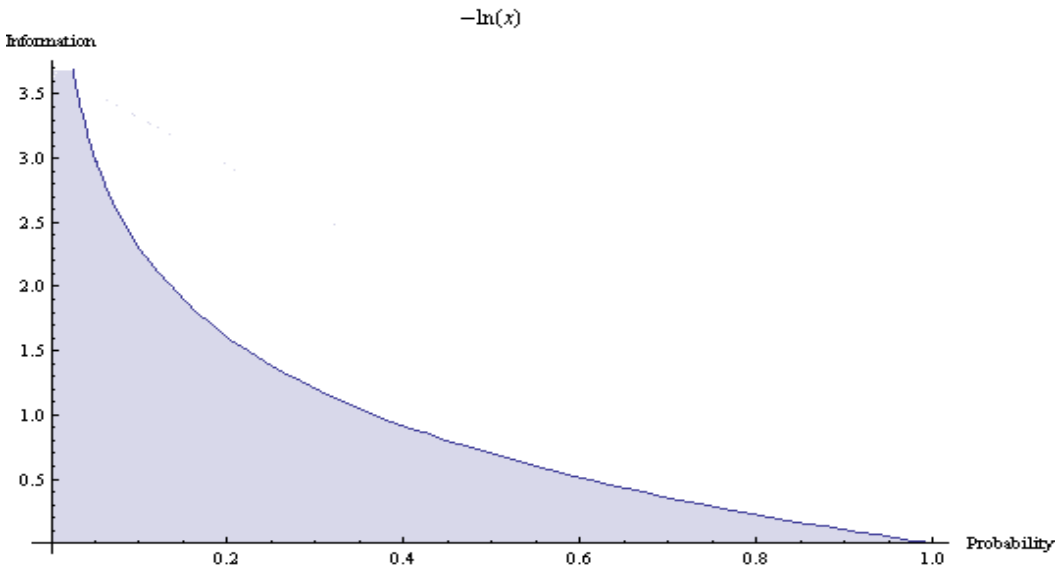

Figure 2 Quantity of information for a given probability

There are three types of missing values: completely at random, at random, and not at random; the first two types are of no value because one cannot exploit randomness. Information in non-trivial structure (i.e., missing not at random) can possibly be used to help in classification if no imputation was made. Research in machine learning has demonstrated that missing values can be exploited without imputation [Chechik et al, 2007].

2.3.2. Case-based reasoning

Case-based reasoning is a well known way to exploit knowledge. It has been already considered for anomaly detection in the maritime domain [Gupa et al 2009];[Bergeron Guyard et al., 2009]. The main challenge of cased-base reasoning is to find cases similar to the current case in the knowledge base. The most popular way of doing this is to define a sum of weighted distance measures to compare every case. This is easy to set in place and can give good results when fine tuned. However, the subset of cases can be irrelevant because this constant similarity measure may not be optimal over all the knowledge base.

Similarity measures can also be based on classification methods. However, these methods have rigid classification boundaries and are prone to error [Cunningham, 2009]. They are meant to create classes, not to qualify a particular event. As said in [Bergeron Guyard et al., 2009], other approaches to evaluate similarity of situations should also be considered and compared in the context of anomaly detection.

2.3.3. Reasoning with limited resources

Anomaly detection systems should give the maximum support they can to operators. In the maritime domain, the objective is to be able to classify vessels with almost no prior knowledge [DARPA, 2005]. However, most current methods are based on data density and thus, need a lot of data. In low density regions, anomaly detection is either not performed at all or an alert is

automatically raised. In both cases, systems perform poorly, leaving the operator on his/her own. Here are some papers where these problems can be seen : [Baldacci, 2008];[Baldacci, Carthel, 2009];[Bomberger et al, 2006];[Guerriero et al, 2010];[Rhodes et al, 2007];[Ristic et al, 2008];[Riveiro et al, 2008].

3. Iterative sub-setting

The work presented here addresses the previous issues. Instead of trying to apply an existing solution to these problems, a novel method has been created. All the work presented here is experimental and the main goal was to provide a proof of concept. Taking this into account, one should not expect that this presentation covers in details every technical or theoretical aspects. To achieve a state were preliminary results can be provided, some naïve technical methods were used. These methods are identified and will have to be revisited in future work.

The result can be described as a CBR system with an information theoretic similarity measure. Distributions are computed over the selected case similar to a new observation. The resulting distributions are then used to draw conclusions or used to perform inference. Iterative Sub-Setting (ISS) is similar to decision tree construction schemes ID3 and C4.5 presented in [Quinlan, 1993] and the JAVA implementation of C4.5: J48. For this reason, comparison will often be made between these techniques in the rest of this document.

The ultimate objective of this work is to provide an alternative solution to the current CBR technology used in the Multi-Intelligence Tools Suite developed at DRDC Valcartier.

3.1. Data characterization

The repository used to perform case-based reasoning consists in a list of multidimensional feature vectors. The reasons for this choice are that the repository will evolve over time and also that it can contain missing values. In contrast, classical OnLine Analytical Processing (OLAP) systems use a static hyper-cube where data cleansing and data reduction methods have been applied.

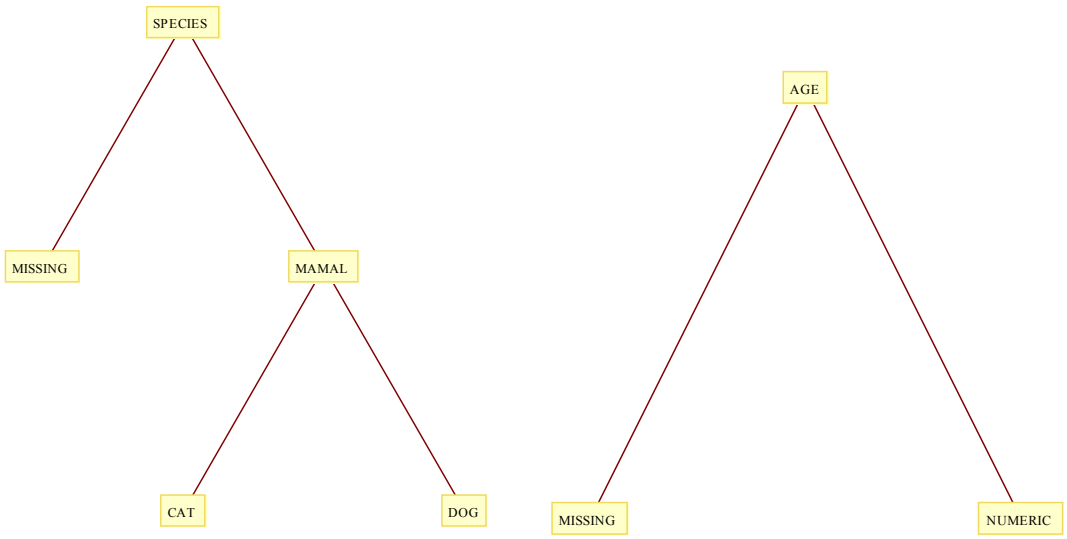

Figure 3 Examples of context hierarchy with missing values

All features are accessed using a concept hierarchy. The first level is used to handle missing values by considering features as categorical. It splits the data into exiting values and missing ones. For example, if the feature is the speed of vessels, one category would be numerical values and the other would be the missing ones. The category representing non-missing values can be expanded (but this is not mandatory) with a classical context hierarchy as shown in [Han, Kambert, 2001] and suggested in [Quinlan, 1993]. Figure 3 pictures this example.

3.2. Similar cases selection

The inference of distributions is done using subsets of the repository. These subsets are used for case-based reasoning and they are selected using a modified ID3 algorithm. The first difference from ID3 is that instead of creating the whole tree, a branch is drilled-down for each new event. The path of the branch is based on the information gain calculation using the Kullback-Leibler divergence. The second difference is that the drill-down process follows the context hierarchy. Finally, while numerical values need to be discretized in ID3 or split in two for C4.5, ISS takes a subset of the data surrounding the feature vector value. This process is done iteratively, just like in C4.5. However, while C4.5 finds the optimal splitting point (from a classification perspective), ISS only weeds out extreme values, i.e., values far from the one in the feature vector. For now, extreme values are defined as the one being at more than one time the standard deviation (from the feature distribution) of the event feature value. This sub-optimal scheme should be revisited. The process stops when a subset reaches the desired number of cases or when no more cases can be removed.

This process provides solutions for some issues presented earlier. Information in missing values will be used if a class of missing values provides a distinctive distribution. A list of cases is always provided to try to explain an event. Selected cases are contextually relevant because they surround the event feature vectors and they are selected to try to maximize the information gain. The highest level of abstraction is always used as a consequence of the top-down hierarchy visiting scheme. Handling time is relatively easy using a carefully designed time hierarchy. Unlike similarity measures based on weighed distance, ISS does not need to be tuned to work on specific primitive data types. It can work efficiently even with no hierarchy. Finally, there are no rigid classification boundaries like the ones in C4.5 because of the surrounding subset. Figure 4 gives a visual example of this problem.

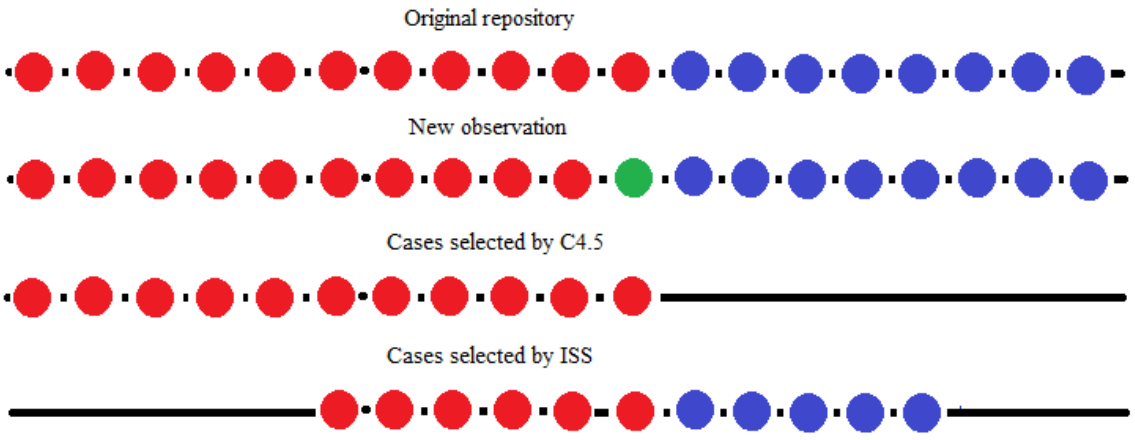

Figure 4 Difference in selected cases between C4.5 and ISS

3.3. Cases repository evolution

There are many ways to make the repository evolve over time. Adding cases is obviously a simple way to increase the coverage of the repository. Removing obsolete cases prevent the repository to become too large. Usage statistics, prevention of duplicate entries and coverage are factors one should consider when updating the repository. Details of these considerations are presented in [Watson, 1998]. The two implementations used here are much simpler. The first one consists in a fixed size repository where the first in is the first out. The second inserts new cases in front of the list and removes cases at the end. However, it moves every cases used during a classification in front of the list. These two schemes are evaluated below.

3.4. Anomaly detection

There are two ways of performing anomaly detection. The first step is to infer the values of interest. In the case of labelled anomalies, the distribution of the labels will be inferred. The new event will be classified upon the most probable label. In case of unlabelled data sets, distributions are inferred for each feature. Values from the new event are tested to see if they belong to these distributions (using Grub's test for example). If not, an anomaly is detected.

3.5. Speed issue

By definition, ISS is an iterative process that removes irrelevant cases from the case base. This process is time consuming. The reason is that for all iterations, a statistical distribution of all features must be computed. Also, subsets must be extracted from the main case base for every iteration. The complexity of the classification procedure appears to be (this is not proven yet) O(m n log n) where m is the number of attributes and n is the number of cases in the case base. This complexity is the same as the initial tree building of C4.5, However, ISS does not compute the whole tree to classify a new case. This limitation can be a major issue for a large case base or for the real-time classification of multiple observations.

4. Performance result

This section provides results to support some of the claimed benefits of ISS. Four aspects have been tested: confidence level, evolution, impact of repository size and the impact of data cleansing. Other aspects of ISS, like the drill down capability, are not quantifiable and thus have not been tested for this technical note. The testing procedure takes a subset of the main dataset as a case base and uses it to predict the remaining cases. This procedure is performed at least 30 times with different subsets for every test.

The proof of concept was done on home equity loan cases taken from the SAS enterprise miner. The objective of the test is to determine who should be approved for a home equity loan. The target variable is a binary variable that indicates whether an applicant eventually defaulted on the loan. The input variables include the amount of the loan, the amount due on the existing mortgage, the value of the property, and the number of recent credit inquiries.

Of course all the results from these tests depend on the capability of the dataset to provide enough information to infer the binary variable. Like any other data mining algorithm, ISS cannot infer correct results when the predicted variable is uncorrelated to other variables in the dataset. This is not a problem with this dataset.

4.1. Confidence level test

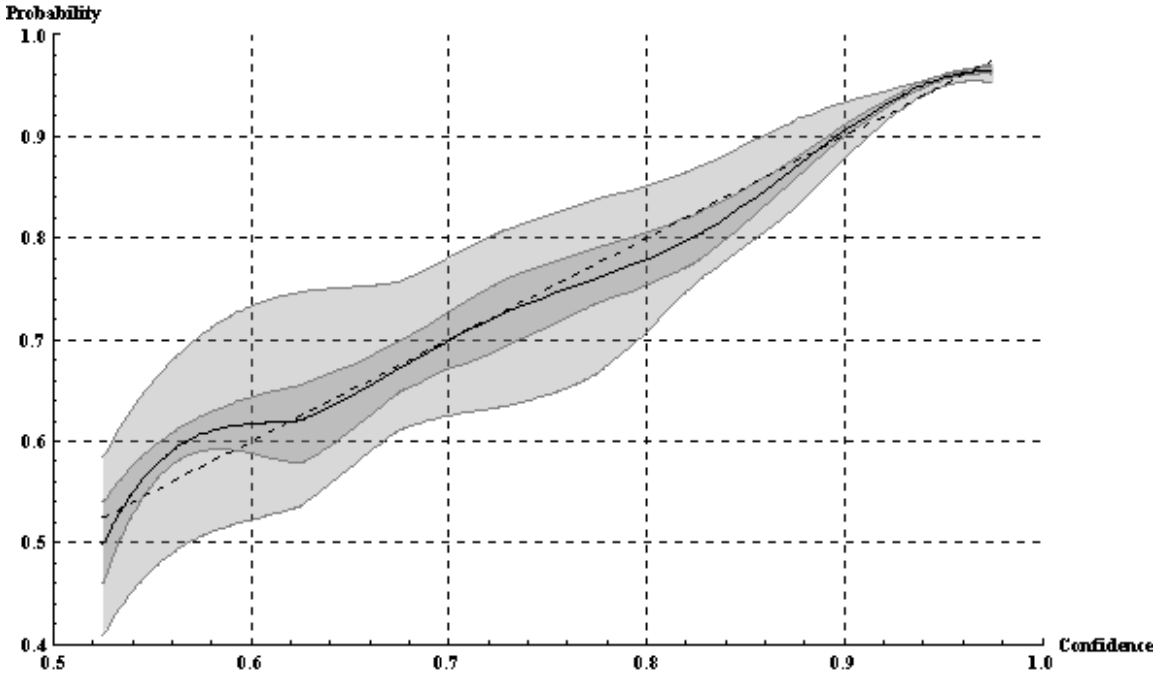

Figure 5 Confidence level

If the ISS procedure performs a classification on labelled data and gives a confidence level, one would want to know if he/she can actually trust this number. Multiple runs of the procedure were made and Figure 5 shows the results. The black dashed line is the perfect case, the light gray regions cover all the runs, the dark gray shows only results between the first and the third quartile and the black line is the average. One can observe that the confidence level matches closely to the actual probability. Knowing that sixty percent of the predictions have a confidence level above 0.9, the average expected error on the confidence level is around 5 percent.

4.2. Evolution test

To simulate an evolution in the dataset, all the cases were sorted in increasing order of loan amount. The initial case base was taken at the bottom of that list and predicted cases were fed in increasing order. Figure 6 presents the density distribution of confidence values. The black dashed line is the static case base, the dotted line represents the move to front scheme, and the solid gray line is the First In First Out (FIFO) scheme. It is clear that the two evolution methods greatly outperform the static model.

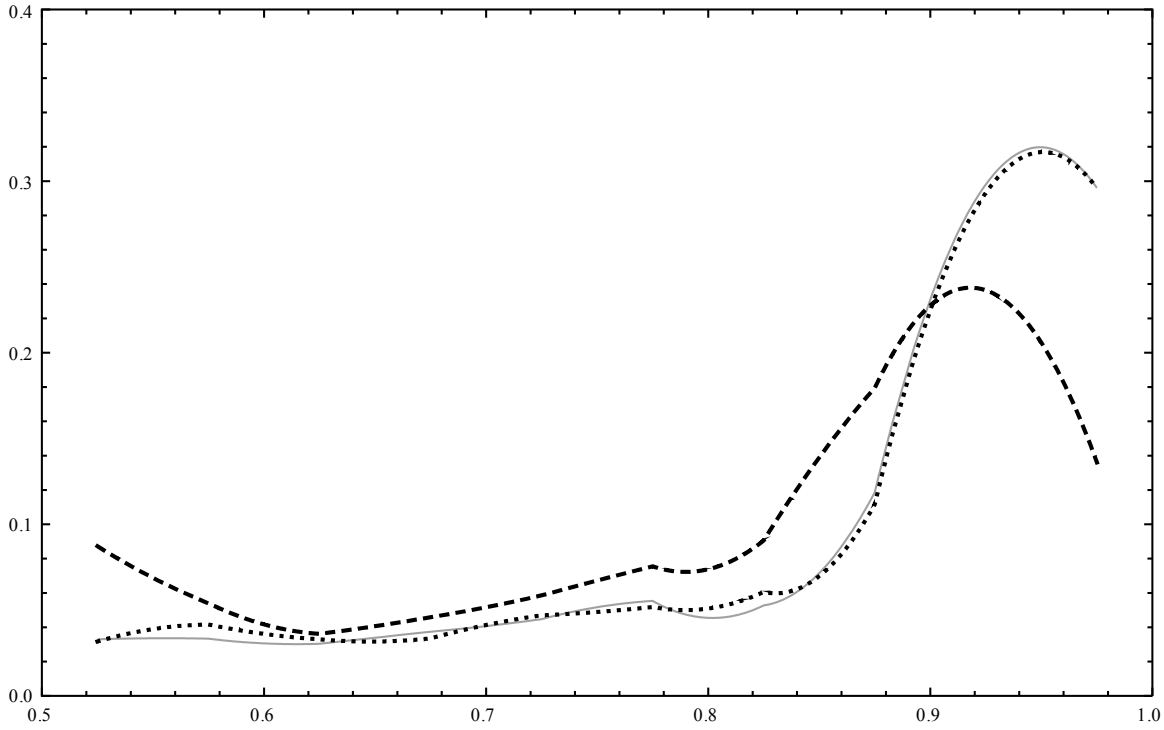

Figure 6 Evolution test density distributions of confidence level

4.3. Repository size test

The size of the case base has a major incidence on the confidence level density distribution. In this dataset, adding more cases to the case base reduces the uncertainty. In Figure 7, the gray line

has fifty percent of the dataset in the case base while the black dashed line as only five percent. ISS still perform well with minimal number of cases. In the case of a large case base, the confidence level density is moved toward a greater certainty.

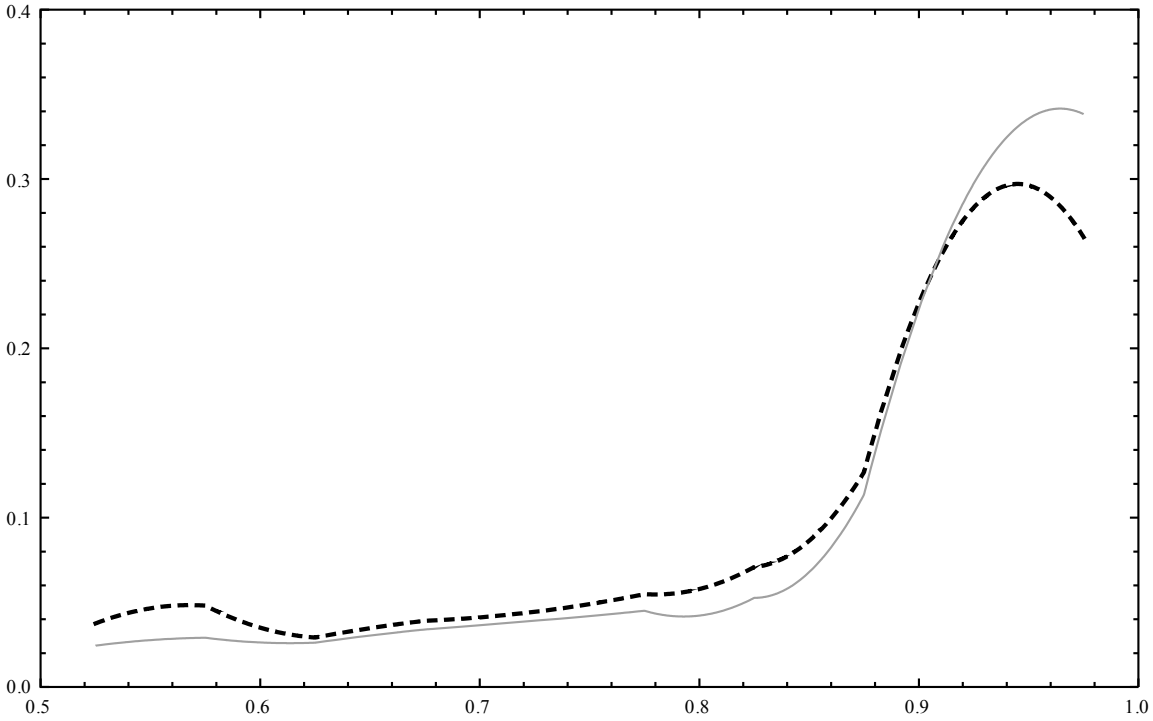

Figure 7 Repository size test density distributions of confidence level

4.4. Impact of data cleansing

In this last test, an evaluation of the impact of imputation is performed. In Figure 8, the dashed line represents the performance on the original dataset. The dotted line represents the performance on a dataset where the imputation was made with ISS. Finally, the solid line represents the performance on a dataset where the imputation was done by replacing missing value with the mean. These results show that instead of hurting the performance, missing values provide valuable information and increase certainty. Note that ISS was not tested against more statistically advanced techniques.

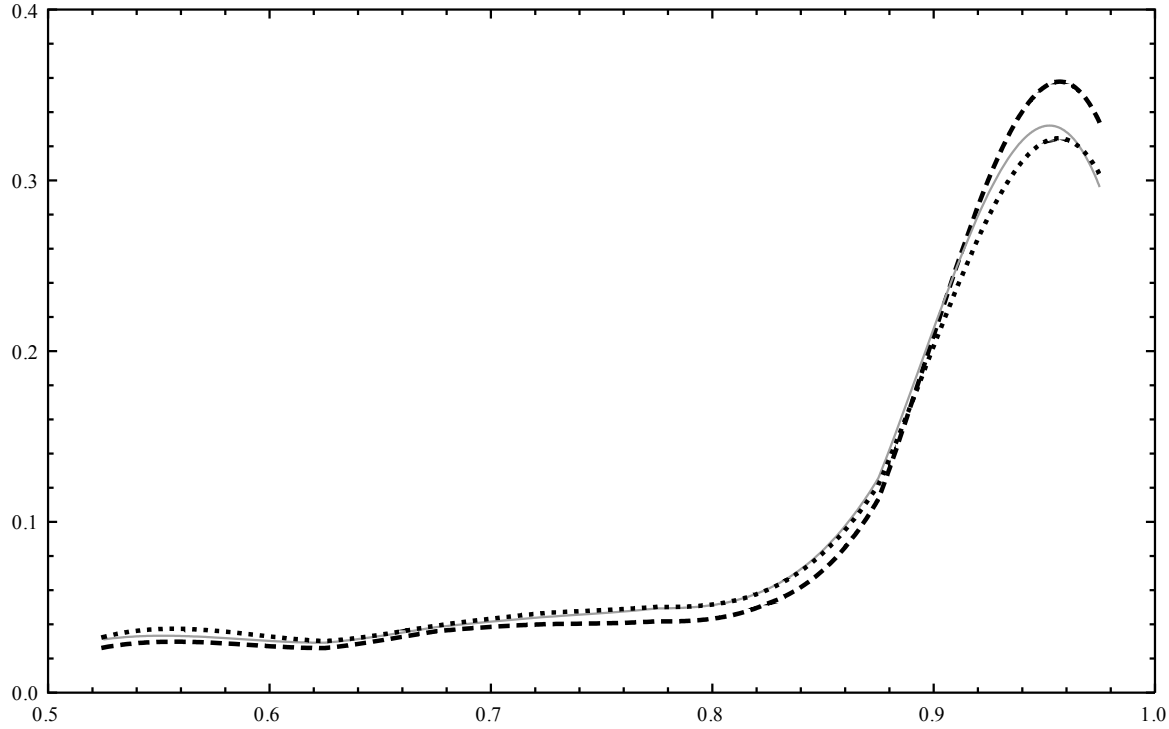

Figure 8 Impact of data cleansing density distributions of confidence level

5. Future work and recommandations

The work done so far was just a proof of concept. The goal was to test and assess new ideas to see if there is an opportunity to improve the MITS platform. Many aspects of the depicted algorithm of ISS can be (and must be) improved before it is integrated in any existing products. This section provides an overview of the remaining work to be done and formulates some recommendations.

5.1. Work to be done

One aspect that needs further research is the subset selection for real values. C4.5 selects a splitting point that maximizes the information gain. ISS should search for the optimal subset when it iterates instead of selecting an arbitrary one. It is also not known if this optimal selection over real values during the iteration process will lead to an optimal final subset. For this reason, a complete comparison of the performance of C4.5 and ISS should be done on reference datasets.

The complexity of ISS appears to be O(m n log n) where m is the number of attribute and n is the number of cases in the case base. The computation time can be easily reduced by performing the calculation on many different machines. Using a cluster, the complexity can be reduced to a maximum of O(n log n) which is a reasonable improvement. The ISS procedure should be adapted to exploit this possibility.

The case base consists of an ordered list of cases. The procedure must go through all cases to evaluate distributions and to extract subsets. Another potential avenue to improve the speed of ISS could be to store the cases in a graph. Each case would be connected to other cases using attribute affinity. The affinity measure should be based on information theory to link cases with attributes that behave in the same manner. Methods in network analysis to find clique, clubs and clan could be used to achieve this.

5.2. Recommendations

ISS is still a proof of concept. However, some recommendations can be made to improve the current MITS platform. In order to be able to be used with different types of data, the similarity measure used should be generic. Information theoretic measures can handle directly all type of primitive data. Weighted distance similarity measures are specific to the data and must be changed for each new data set. For this reason, the used of ISS, C4.5 or J48 would give the MITS more flexibility. Moreover, since ISS addresses many current problems in MAD, efforts should be put to perform a more complete evaluation of this technology. If the results are up the expectation, ISS would be a good asset in the MITS.

6. Conclusion

This document presented a novel method to perform CBR. The motivation for this work was first presented to highlight gaps sought to be filled. The ISS algorithm was then introduced with some empirical results to demonstrate the potential benefits of this method. From these results, recommendations were made to improve the current CBR capability of the MITS platform, i.e., adding an information theoretic similarity measure to the current weighted distance similarity measure.

Moreover, future research directions to complete this work were identified; much is left to be done. The ISS method presents potential improvements over current technologies, and efforts need to be invested to complete the validation of these valuable improvements.

Finally, ISS is more than just a CBR similarity measure; it is data-mining tool. It can be used in a variety of situations because of its flexibility and its robustness to missing data. Beyond the sought use of ISS in CBR or MAD, it can be used for anomaly detection in general or as a network mining tools. ISS is a good research opportunity on its own, and it opens new potential avenues of research. Considerations leading to the creation of ISS raise questions on the utility of some accepted techniques. Is it necessary to perform imputation, PCA or modelling? These questions rose by ISS need to be investigated in future research.

References

[Baldacci, 2008], Baldacci, A., *AIS Emission Anomaly Detection in Support of Maritime Surveillance*, Technical Report, NURC-FR-2008-020, July 2008.

[Baldacci, Carthel, 2009], Baldacci, A. and Carthel, C., *Maritime Ttraffic Characterization with the Automated Identification System*, Technical Report, NURC-FR-2009-008, May 2009.

[Bergeron Guyard et al., 2009], Bergeron Guyard, A. and Roy, J., *Towards Case-Based Reasoning for Maritime Anomaly Detection*, Second IEEE Symposium on Computational Intelligence for Security and Defense Applications - Detecting and Adapting to Emerging Threats, Ottawa, Canada, 8-10 July 2009

[Berry, Linoff, 1997], Berry, M. J. A. and Linoff, G. (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Support*. New York: John Wiley&Sons.

[Bomberger et al, 2006], Bomberger, N.A., Rhodes, B.J., Seibert, M. and Waxman, A.M., *Associative Learning of Vessel Motion Patterns for Maritime Situation Awareness*, In Proceedings of the 9th International Conference on Information Fusion (Fusion 2006), Florence, Italy, July 10-13, 2006.

[Chechik et al, 2007], Chechik, G., Heitz, G., Elidan, G., Abbeel, P., and Koller, D. (2007). *Max-Margin Classification of Incomplete Data*. Advances in Neural Information Processing Systems 19.

[Cunningham, 2009], Cunningham P., A Taxonomy of Similarity Mechanisms for Case-Based Reasoning, IEEE Transactions on Knowledge and data Engineering, vol. 21, no. 11, November 2009.

[DARPA, 2005], DARPA, *Proposal Information Package (PIP) - Predictive Analysis for Naval Deployment Activities (PANDA) - BAA 05-44*, 28 September 2005.

[Guerriero et al, 2010], Guerriero M., , Coraluppi S. and Carthel C., *Analysis of AIS Intermittency and Vessel Characterization Using a Hidden Markov Model*, Technical Report, NURC-FR-2010-002, January 2010.

[Gupa et al, 2009], Gupta K.M., Aha, D.W. and Moore P.G., *Case-based Collective Inference for Maritime Object Classification*. Manuscript submitted for review (2009)

[Han, Kambert, 2001], Han J. and Kambert M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2001.

[Hutchins et al, 2009], Hutchins S. G., MacKinnon D. J., Freeman J., Gallup S. P., *Maritime Domain Awareness: Assessment of Current Status*, In Proceedings of the 14th International Command and Control Research and Technology Symposium (ICCRTS), Washington, DC. June 15-17, 2009.

[Quinlan, 1993], Quinlan, J. R., C4.5: *Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, Ca, 1993.

[Rhodes et al, 2007], Rhodes, B.J., Bomberger, N.A. and Zandipour, M., *Probabilistic Associative Learning of Vessel Motion Patterns at Multiple Spatial Scales for Maritime Situation Awareness*, Proceedings of the 10th International Conference on Information Fusion (Fusion 2007), Quebec, Canada, 9-12 July 2007.

[Ristic et al, 2008], Ristic, B., La Scala, B., Morelande, M. and Gordon, N., *Statistical Analysis of Motion Patterns in AIS Data - Anomaly Detection and Motion Prediction*, in Proceedings of The 11th International Conference on Information Fusion (Fusion 2008), Cologne, Germany, June 30 – July 03, 2008.

[Riveiro et al, 2008], Riveiro, M., Falkman, G. and Ziemke, T., *Improving Maritime Anomaly Detection and Situation Awareness Through Interactive Visualization*, Proceedings of the 11th International Conference on Information Fusion (Fusion 2008), Cologne, Germany, 30 June - 03 July, 2008.

[Seibert, 2009], Seibert, M., *Maritime Anomaly Detection*, Workshop on Detection of Anomalous Behaviors in Maritime Environments, Carnegie Mellon University, 25-26 June 2009.

[Watson, 1998], Watson I., *Applying Case-based Reasoning: Techniques for Enterprise Systems*, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1998

This page intentionally left blank.

List of symbols/abbreviations/acronyms/initialisms

CBR	Case-Based Reasoning
DAM	Détection d'Anomalies Maritimes
FIFO	First In First Out
ISS	Iterative Sub-Setting
MAD	Maritime Anomaly Detection
MITS	Multi-Intelligence Tools Suite
OLAP	OnLine Analytical Processing
PCA	Principal Component Analysis
R&D	Research & Development

This page intentionally left blank.

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

Defence R&D Canada – Valcartier
2459 Pie-XI Blvd North
Quebec (Quebec)
G3J 1X5 Canada

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

UNCLASSIFIED

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

Iterative sub-setting

4. AUTHORS (last name, followed by initials – ranks, titles, etc. not to be used)

Martineau, E.

5. DATE OF PUBLICATION (Month and year of publication of document.)	6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)	6b. NO. OF REFS (Total cited in document.)
December 2010	30	18

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Technical Memorandum

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

Defence R&D Canada – Valcartier
2459 Pie-XI Blvd North
Quebec (Quebec)
G3J 1X5 Canada

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC Valcartier TM 2010-461	

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

Unlimited

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

Unlimited

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

The objective of this document is to communicate the progress made on an initiative to improve the current Case-Based Reasoning (CBR) functionality of the Multi-Intelligence Tools Suite (MITS) platform. To do so, a new similarity measure based on information theory called Iterative Sub-Setting (ISS) is proposed. The goal of this initiative was to provide a potential solution to a list of issues identified in the literature about Maritime Anomaly Detection (MAD). Various tests with this measure were made to demonstrate the potential for it to address selected issues in MAD. Results show that this method is more sensible to the context where the CBR is performed, can be used with a small case base, evolve over time and provide an accurate confidence level on the selected cases.

L'objectif de ce document est de communiquer les progrès réalisés grâce à une initiative pour améliorer la capacité actuelle de raisonnement à base de cas de la plate-forme « Multi-Intelligence Tools Suite (MITS) ». Le but de cette initiative était de fournir une solution potentielle à une liste de problèmes identifiées dans la littérature sur la détection d`anomalies maritimes (DAM). On a proposé une nouvelle mesure de similarité basée sur la théorie de l'information. Différents tests ont été effectués avec cette mesure pour démontrer son potentiel à résoudre certaines problèmes en DAM. Les résultats montrent que cette méthode est plus sensible au contexte dans lequel le raisonnement à base de cas est effectué; elle peut être utilisée avec une base de cas limitée, évoluer dans le temps et fournir un niveau de confiance précis sur la pertinence des cas sélectionnés.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Case-Based Reasonning Similarity measure, Data-Mining, Anomaly detection

Defence R&D Canada

Canada's Leader in Defence
and National Security
Science and Technology

R & D pour la défense Canada

Chef de file au Canada en matière
de science et de technologie pour
la défense et la sécurité nationale

www.drdc-rddc.gc.ca